Un système hybride de recherche d'information intégrant le raisonnement à partir de cas et la composition d'ontologies

Ghada Besbes*, Hajer Baazaoui-Zghal*
Henda Ben Ghezela*

*Laboratoire Riadi-GDL, ENSI Campus Universitaire de la Manouba, Tunis, Tunisie
ghada.besbes@gmail.com, hajer.baazaouizghal@riadi.rnu.tn, henda.bg@cck.rnu.tn

Résumé. La croissance des informations disponibles sur le web nécessite des outils de recherche de plus en plus performants permettant de répondre efficacement aux besoins des utilisateurs. Dans ce contexte, l'utilisation des ontologies présente des atouts importants. Cependant, la construction manuelle d'ontologies est très coûteuse, ceci a poussé à proposer des approches permettant d'automatiser cette construction. Cet article présente un système de recherche d'information hybride basée sur le Raisonnement à Partir de Cas (RàPC) et la composition d'ontologies. Ce système vise à combiner la construction automatique d'ontologies modulaires et le RàPC, qui a pour but d'améliorer les résultats de recherche d'information (RI). Des expérimentations ont été menées et les résultats obtenus montrent une amélioration de la précision dans le cas d'une recherche d'information sur le Web.

1 Introduction

Les ontologies ont contribué au succès des moteurs de recherche sémantique et sont de plus en plus utilisées pour améliorer la recherche d'information sur le web et la reformulation des requêtes. Cependant, la construction d'ontologies est généralement un processus long et coûteux, le recourt aux ontologies modulaires constitue une piste prometteuse. Une ontologie modulaire est une ontologie qui référence un fragment d'une ontologie de domaine et a l'avantage d'être réutilisée ultérieurement. La composition d'ontologies permet de construire une ontologie modulaire à partir d'un ensemble de modules ontologiques qui constituent un réseau, et permet d'améliorer l'organisation des concepts sémantiques.

Un état de l'art a permis de constater que les approches de composition d'ontologie proposées ne considèrent pas les relations sémantiques entre les termes, elles ne sont donc pas expressives et ne peuvent pas être efficacement utilisées pour la recherche sémantique sur le Web. Nos travaux précédents ont proposé en un premier lieu une approche de recherche d'information basée sur le RàPC pour reformuler la requête de l'utilisateur et lui recommander des résultats fondés sur les cas stockés (Elloumi-Chaabene et al., 2010), et par la suite une nouvelle méthode de composition de modules ontologiques basée sur des mesures de similarité sémantique (Elloumi-Chaabene et al., 2011). L'objectif du présent travail est de proposer un système hybride de recherche d'information basée sur le RàPC et la composition d'ontologies, ayant

pour but d'améliorer la précision des résultats fournis aux utilisateurs et de répondre à ses besoins. Notre contribution porte sur l'intégration de : 1/la composition de modules ontologiques pour construire une ontologie modulaire ce qui améliore le processus d'enrichissement de requête ; 2/ RàPC pour prendre en considération les préférences de l'utilisateur ; 3/une base de connaissances (BC) qui prend en considération le contenu des documents Web jugés pertinents par l'utilisateur lors des recherches précédentes afin d'enrichir la requête. Notre motivation est d'utiliser des requêtes passées afin d'améliorer la précision des résultats fournis aux utilisateurs et d'utiliser aussi les informations dans les documents Web pertinents pour enrichir la requête de l'utilisateur.

Ce papier est organisé comme suit : la section 2 introduit notre système qui intègre la composition d'ontologie et le RàPC, que nous détaillons dans les sous sections. Dans la section 3, nous présentons les expérimentations menées. La section 4 conclut et aborde nos perspectives.

2 Un système hybride de recherche d'information basé sur le raisonnement à partir de cas et la composition d'ontologies

Le système proposé se compose de quatre composants principaux : (1) un composant pour le RàPC (2) un composant pour la composition d'ontologies (3) un composant pour la base de connaissances et (4) un composant pour la classification de documents. L'architecture générale du système est présentée par la figure 1. Dans les sous-sections suivantes nous détaillons les différents éléments du système proposé.

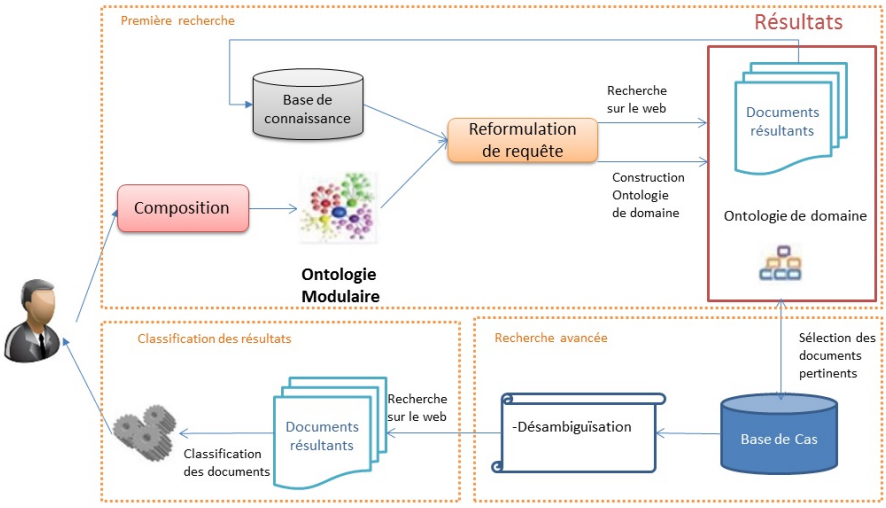

FIG. 1 – *Architecture générale du système.*

2.1 Première recherche

La première recherche sur le Web offre une reformulation de la requête initiale de l'utilisateur. En effet, le processus de recherche se déroule comme suit : l'utilisateur sélectionne les modules ontologiques nécessaires pour la composition, pose sa requête, qu'on enrichit par l'ontologie modulaire et la BC. En effet, la similarité sémantique est calculée entre les concepts de l'ontologie et les termes de la requête. Les concepts les plus similaires à la requête sont utilisés pour l'enrichir. La mesure de similarité utilisée est PMI_IR (Pointwise Mutual Information). Elle a été adaptée au Web par (Turney, 2001) en définissant p(a) la probabilité du terme " a " dans le web. Cette probabilité est estimée à partir du nombre de résultats retournés en cherchant le terme a sur le web. La mesure PMI_IR est calculée selon (1)

$$PMI\_IR = \log_2 \frac{p(aANDb)}{p(a)\times p(b)} \quad (1)$$

Un enrichissement par la BC est ensuite effectué. L'utilisateur valide la requête enrichie et il choisit le domaine de recherche. Une recherche sur le web commence en même temps que la construction de l'ontologie de domaine choisi. Le résultat de cette première recherche est donc un ensemble de documents et une ontologie de domaine qui vont permettre de passer à la recherche avancée.

La composition d'ontologies. La composition d'ontologies vise la construction d'une ontologie modulaire en utilisant un ensemble de modules ontologiques ceci permet d'améliorer leur structure en considérant les relations taxonomiques et non-taxonomiques implicites entre les concepts. Elle est évaluée par des mesures de cooccurrence basées sur le web. Les principales étapes de la méthode proposée sont les suivantes : (1) La réorganisation des modules ontologiques, qui consiste à déterminer les modules qui ont des concepts en communs et collecter l'ensemble des concepts qui se chevauchent pour chaque paire de modules. (2) La classification des concepts obtenus en modules en fonction de leur similarité sémantique en utilisant les mesures de cooccurrence (Turney, 2001). L'idée principale est de construire un graphe de cooccurrence et d'appliquer un algorithme de clustering pour réorganiser les concepts et déterminer l'ensemble des nouveaux modules. (3) La construction d'une structure hiérarchique des modules ontologiques qui constituent l'ontologie modulaire.

La base de connaissances. Ce composant gère la BC composée d'une base de faits, une base de règles et un moteur d'inférence. Le système remplit automatiquement les règles de cette base par les termes les plus fréquents des documents jugés pertinents par l'utilisateur. Elle est utilisée pour enrichir la requête de l'utilisateur. En effet, suite au raisonnement fait par le moteur d'inférence sur les règles et les faits de la base, de nouvelles conclusions sont obtenues. Ces conclusions sont les termes significatifs extraits des documents pertinents de recherches antérieures qui vont être ajoutées à la requête pour l'enrichir. Les règles de cette base sont des règles de la logique d'ordre zéro écrites sous la forme :

$Request -> BestTerm1,,,BestTermi,,,BestTermN.$

Avec N le nombre de documents choisis et BestTerm i le terme ayant la plus haute fréquence dans le document i. Lorsqu'une nouvelle requête est soumise au système, le moteur d'inférence utilise ces mots clés pour faire un raisonnement : 1/ si la requête existe, il renvoie

les termes significatifs extraits de documents pertinents retournés lors d'une recherche précédente pour cette requête (nouvelles conclusions) 2/ si la requête n'existe pas, la reformulation se fera seulement par l'ontologie modulaire.

2.2 Recherche avancée

L'utilisateur sélectionne les documents pertinents à partir de ceux récupérés dans la première recherche. Ensuite, le système insère d'une part un nouveau cas et sa solution, et d'autre part, utilise des techniques de fouille de texte pour extraire les termes les plus fréquents et les ajouter à la BC. A partir de l'ontologie du domaine construite, l'utilisateur choisit le concept de recherche. En utilisant WordNet (Miller, 1990), les synonymes, hyponymes et hyperonymes sont insérés dans la BDC pour mettre à jour la signature sémantique du cas(c'est une liste de termes qui apparaissent fréquemment avec le concept du cas). En se basant sur le concept choisi et la BDC, une nouvelle recherche est possible et comprend : 1/ une désambiguïsation : le système offre les sens du concept choisi à partir de WordNet et une recommandation basée sur un algorithme de désambiguïsation sémantique ; 2/ l'ajout de termes : l'utilisateur ajoute des concepts de son choix afin d'enrichir l'ontologie ; 3/ les recommandations : ce sont les cas similaires de recherche (des URLs recommandés qui correspondent à des cas pertinents de recherche similaires stockés dans la BDC).

Le raisonnement à partir de cas. La combinaison des ontologies et le mécanisme de RàPC peut améliorer les performances des recherches sur le Web sémantique. Le but ici est d'enrichir automatiquement la requête à l'aide de requêtes antérieures effectuées par l'utilisateur. Ce composant gère la BDC, en adoptant le modèle vectoriel pour représenter le cas. Typiquement un cas contient au moins deux parties : une description de situation représentant un "problème" et une "solution" utilisée pour remédier à cette situation. Le problème comprend le domaine général, le concept spécifique de la recherche ainsi que la signature sémantique. La solution correspondante est composée des documents jugés pertinents par l'utilisateur ainsi que deux vecteurs, domaine et module, associés aux concepts du domaine et concept de recherche respectivement. Ces deux vecteurs correspondent aux poids des concepts du domaine et de la signature sémantique calculés par la mesure TF*IDF dans les documents résultats. A l'aide de la BDC, on peut insérer un nouveau cas, le mettre à jour ou bien rechercher des cas similaires.

2.3 Classification des résultats

A partir des choix effectués par l'utilisateur au cours de l'étape précédente (la désambiguïsation etc.) ainsi que les requêtes similaires dans la BDC, le système reformule la requête et recherche de nouveaux résultats sur le web. Les documents récupérés sont classés par ordre de pertinence par rapport à la requête et ceux sélectionnés par l'utilisateur sont ajoutés à la BDC et leurs termes les plus fréquents sont ajoutés à la BC.

La classification des documents. Afin de récupérer les documents les plus pertinents, le modèle de Salton (Salton et Rocchio, 1983) est utilisé (les termes remplacés par des concepts). Un filtrage par les vecteurs domaine et module est appliqué pour ne garder que les documents dans la même thématique et appartenant au même module. Chaque document est représenté

par un vecteur Dj = (d1j ; d2j ; ... ; dnj). Où dij est le poids du concept ci dans le document Dj , N étant le nombre de concepts dans la signature sémantique. Le vecteur Q = (Q1 ;Q2 ; ... ;Qn) représente la requête, où Qi est le poids du concept ci dans la requête. La mesure de similarité entre un document et une requête est calculée avec la formule cosinus :

$$Sim(D_j; Q) = \frac{\sum_{i=1}^{N} d_{ij} \times q_i}{\sqrt{\sum_{i=1}^{N} d_{ij}^2 \times \sum_{i=1}^{N} q_i^2}}$$

3 Expérimentation

Un prototype supportant le système proposé a été développé pour fournir une interface utilisateur qui permet la manipulation des ontologies, la gestion de la BC et l'affichage des résultats à partir du moteur de recherche sur le web (en utilisant l'API de Bing). L'évaluation expérimentale des performances du système proposé est menée en comparant les résultats de la précision moyenne des différents systèmes. Quatre scénarios ont été testés : 1/ une recherche classique (c'est une recherche par mots clés sans effectuer une reformulation de la requête), 2/ une reformulation de requêtes en utilisant l'approche de composition d'ontologies, 3/une recherche basée sur l'approche RàPC et finalement 4/ une recherche hybride. L'évaluation expérimentale a été conduite en utilisant le SRI expérimental LEMUR [1], largement utilisé par la communauté RI. Les différents tests sont menés sur la collection INEX 2010 [2].

Les résultats illustrés par la figure 2 représentent la précision exacte à 10, 30, 50 et 100 documents, et on observe une amélioration significative de la pertinence de l'information récupérée lors de la reformulation de requête pour le quatrième scénario.

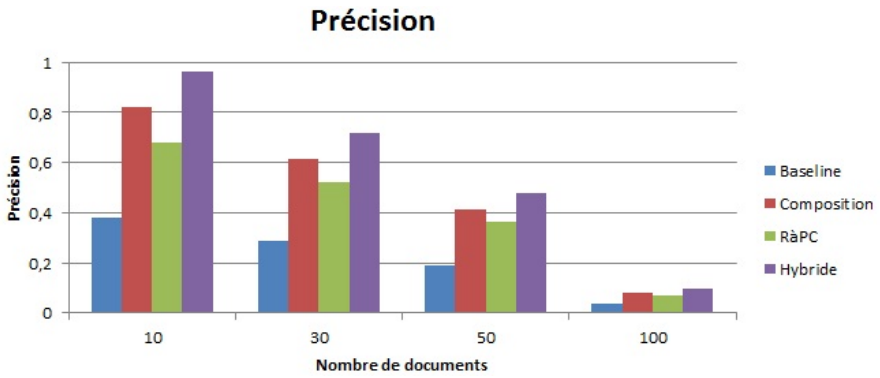

FIG. 2 – *Évaluation et comparaison de la mesure de la précision.*

1. http ://www.lemurproject.org/
2. http ://www.inex.otago.ac.nz/

4 Conclusion

Cet article présente notre système hybride pour la recherche d'information sur le Web intégrant le RàPC et la composition d'ontologies. Tout d'abord, nous nous sommes appuyés sur le RàPC dans le but de prendre en considération les cas déjà rencontrés. Nous avons aussi utilisé une nouvelle méthode de composition des modules ontologiques pour améliorer, d'une part, la structure des modules d'ontologiques et leur organisation interne, et d'autre part, le degré de parenté sémantique entre les concepts et la structure globale de l'ontologie modulaire. Nous avons également intégré une BC pour enrichir la requête de l'utilisateur par les termes les plus fréquents extraits de documents pertinents. Plusieurs techniques de recherche d'information ont été intégrées pour améliorer les résultats de recherche sur le Web. L'expérimentation et l'évaluation menées montrent une amélioration du taux de précision. Dans nos travaux futurs, nous envisageons d'intégrer une nouvelle composante à notre système pour supporter le traitement des systèmes question-réponse afin d'améliorer les résultats de recherche en répondant automatiquement aux questions posées en langage naturel.

Références

Elloumi-Chaabene, M., N. Ben Mustapha, H. Baazaoui Zghal, A. Moreno, et D. Sánchez (2010). Evolutive content-based search system- semantic search system based on case-based-reasoning and ontology enrichment. *In Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, 24–34.

Elloumi-Chaabene, M., N. Ben Mustapha, H. Baazaoui Zghal, A. Moreno, et D. Sánchez (2011). Semantic-based composition of modular ontologies applied to web query reformulation. *In Proceedings of the 6th International Conference on Software and Data Technologies ICSOFT*, 305–308.

Miller, G. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography 3*(4), 235–244.

Turney, P. D. (2001). Mining the web for synonyms : Pmi-ir versus lsa on toefl. *In Proceedings of the twelfth European conference on machine learning*, 491–502.

Summary

The huge number of available documents on the Web makes finding relevant ones challenging. Thus, searching for information becomes more and more complex because of the growing volume of data and of its lack of structure. Ontologies are used to improve the accuracy of information retrieval from the web by incorporating a degree of semantic analysis during the search. However, manual ontology building is time consuming. In this context, this paper suggests a hybrid Information Retrieval System based on Case-Based Reasoning (CBR) and Ontology Composition, aiming to combine automatic modular ontology building and CBR to save previous search queries performed by the user in order to enhance information retrieval results. Some experiments and results obtained with the proposed system are also presented, which show an improvement on the precision of the Web search.